DOCUMENT RESUME

ED 462 437 TM 033 701

AUTHOR Orcutt, Venetia L.
TITLE Computerized Adaptive Testing: Some Issues in Development.
PUB DATE 2002-02-01
NOTE 16p.; Paper presented at the Annual College of Education, University of North Texas Educational Research Exchange (2nd, Denton, TX, February 1, 2002).
PUB TYPE Reports - Descriptive (141) -- Speeches/Meeting Papers (150)
EDRS PRICE MF01/PC01 Plus Postage.
DESCRIPTORS *Adaptive Testing; *Computer Assisted Testing; *Item Response Theory; *Test Construction

ABSTRACT

The emergence of enhanced capabilities in computer technology coupled with the growing body of knowledge regarding item response theory has resulted in the expansion of computerized adaptive test (CAT) utilization in a variety of venues. Newcomers to the field need a more thorough understanding of item response theory (IRT) principles, their impact on CAT development, and other practical issues. This paper provides a brief overview of how a CAT is developed, basic concepts of IRT and proficiency estimation, and examines a few of the issues associated with development to include selected methodologies that address those issues and other methods under investigation. (Contains 19 references.) (Author/SLD)

Running Head: COMPUTERIZED ADAPTIVE TESTING

Computerized Adaptive Testing: Some Issues in Development

Venetia L. Orcutt, M.B.A., P.A.-C.

University of Texas Southwestern Medical Center at Dallas

Paper presented at the annual meeting of the Educational Research Exchange, University of North Texas, February 1, 2002, Denton, Texas.

Abstract

The emergence of enhanced capabilities in computer technology coupled with the growing body of knowledge regarding item response theory has resulted in the expansion of computerized adaptive test (CAT) utilization in a variety of venues. Newcomers to the field need a more thorough understanding of IRT principles, their impact on CAT development, and other practical issues. This paper provides a brief overview of how a CAT is developed, basic concepts of IRT and proficiency estimation, and examines a few of the issues associated with development to include selected methodologies that address those issues, and other methods under investigation.

Computerized Adaptive Testing: Issues in Development

The purpose of testing remains constant regardless of methodology: to compare the proficiency level of one individual or group to that of another for the trait under examination. This may be accomplished by providing the same examination and administration conditions, equating the examinations through common parts, or by connecting the forms through the use of theory. The development of item response theory (IRT) provided the basis for form connection in computerized adaptive testing (CAT). However, until the emergence of enhanced capabilities of microcomputers, CAT remained economically and practically unfeasible. The 1970s and 1980s saw the proliferation of research about various aspects of this testing methodology. This trend has continued as evidenced by the fact that 25% of all paper presentations at the 1999 annual meeting of the National Council on Measurement in Education related to CAT. The body of knowledge surrounding the nuances of CAT has developed more rapidly with the advances in computer technology. With this growth has come new interest from other researchers involved in assessment via rating scales, performance tasks, personality and attitude inventories as well as health outcome measurement. (Hambleton, 1999; Harvey & Hammer, 1999; Hays, Morales, & Reise, 1999; Meijer & Nering, 1999; Wainer, 1983)

As the interest in utilizing CAT has proliferated, newcomers to the field need a more thorough understanding of IRT principles, their impact on CAT development, and other practical issues. Wise (1997) identified a variety of practical issues that arise in the development, implementation, and management of CAT programs. They included item pool development, choice of item response theory models, proficiency estimation method choice, and testing algorithm procedures as well as others. This paper provides a brief

overview of how a CAT is developed, basic concepts of IRT and proficiency estimation, and then will examine a few of the issues associated with development, selected methodologies that address those issues, and other methods under investigation.

Computerized Adaptive Testing Development Overview

The advantage of CAT in providing more efficient and precise ability estimates often sways an organization to convert established paper and pencil or computerized examinations to that of an adaptive test. Development of a CAT follows similar, though more expansive, steps performed in the making of any examination. Once content domains have been agreed upon and test specifications designed, a sufficient number of individual items for each area are constructed through the use of subject matter experts. Item development should include review by sensitivity panels and test development experts. Particular attention must be paid to the variety and difficulty spread of items within each content area. Every new item must undergo pretesting followed by posttest review given the item analysis results provided by both classic test theory and item response theory procedures. Following final item selection, the resulting item pool undergoes simulation studies to assess its function as a pool and to assess the goodness of fit of data to the IRT model selected. Often individual items will undergo revision and retesting prior to the examination becoming "live" (Flaugher, 2000). Overton and Harms (1997) provide an excellent description of many of these steps as they developed a CAT for use in selecting computer programmer trainees used by a large insurance company.

CAT administration is an iterative process of presenting selected items to the examinee and estimating the examinee's level of proficiency given the response to the presented item. The next selected item is matched to the examinee's current proficiency

level and the level is recalculated following the response. The examination continues until a stopping criterion, such as a set level of measurement precision, fixed number of items, or content specifications, is met. (Chen, Ankenmann, & Chang, 2000; Meijer, & Nering, 1999; Thissen & Mislevy, 2000; Wainer & Mislevy, 2000). It is the principles of item response theory (IRT) advanced by Federic Lord, that allow the administration of different sets of items drawn from a pool to different examinees and the estimation of their proficiency on a common scale. The majority of Lord's work targeted dichotomously scored data, however, this work has been expanded to include polytomous and partial credit data. The choice of the IRT model and proficiency estimation method had important implications in the development of a CAT (Hambleton, 2000; Wainer, 1983; Wise, 1997).

Although a more thorough discussion of IRT methods will be presented later, it should be noted that IRT methods are utilized in CAT development, rather than classical test theory, for two main reasons. First, it will allow researchers to more accurately rank respondents in terms of their patter of responses (Crocker & Algina, 1986; Hambleton, 1983). Although some researchers have argued that IRT does not produce scores necessarily different from classical test theory, IRT is maximized at the tails of the distribution (Fan, 1998). This means that CAT will better capture true scores of students who score at extremes of the distribution of scores and potentially discover whether or not students obtained scores accurately or from guessing. Second, using IRT estimates will allow for the generalization of these scores to both the population of interest and to future users, whereas classical test theory results will not generalize to future users.

Basic Concepts of IRT and Proficiency Estimation

Item response theory represents a family of mathematical descriptions utilizing logistic function to depict what occurs when an examinee with particular proficiency level is confronted with a particular test item. At its core is the premise that a single dimension of knowledge or latent trait underlies examinee performance and that all test items rely on this dimension for their correct response. Stated differently, IRT assumes "unidimensionality" that a single ability is responsible for the examinee responses to items. The theory posits various "models" that connect the characteristics or parameters of the test item to the probability of the examinee answering the item correctly. The models differ in the way that proficiency level is presumed to cause the item response given the item characteristics. In the most commonly used models, the examinee proficiency level is denoted as theta (θ) and the item characteristics of discrimination and difficulty level are denoted as a and b respectively with c denoting the impact of guessing. The Rasch model characterizes items with the single parameter of item difficulty while the two-parameter logistic (2-PL) model includes difficulty and discrimination. The three-parameter model adds the c parameter for guessing. It should be noted that additional parameters can be included within the models and such models are the focus of continued research. Although the one parameter IRT model (maximum likelihood estimation) and the Rasch model (unconditional) are only concerned with obtaining estimates for the b parameter, or item difficulty, the estimation methods by which these parameters are calculated differ. Mathematically, the 3-PL model is defined as

$$P(\theta) = c + \frac{(1 - c) \exp [a (\theta - b)]}{1 + \exp [a (\theta - b)]}.$$

The 1-PL model is obtained by fixing $a = 1$ and $c = 0$. The 2-PL model is obtained by fixing $c = 0$ and allowing for the estimation of the a and b parameters (Crocker & Algina, 1986; Hambleton, 2000; Harvey & Hammer, 1999; Wainer & Mislevy, 2000). The resulting graphic representation of this mathematical description of the relationship between the proficiency level and that of the item parameters is known as the item characteristic curve (ICC).

Estimating the proficiency level (θ) can be accomplished through a variety of methods. Two commonly used procedures include the maximum likelihood estimation (MLE) and Bayes Modal Estimate. These methods utilize the probability of an item response pattern following a jth item administered in a CAT and can be represented by the following equation provided in Wainer and Mislevy (2000)

$$P(\mathrm{x_i}|\theta_i,ß) = \Pi\ P_j(\theta)^{Xij} Q_j(\theta)^{1-Xij}$$

where X_i represents the score pattern for examinee i, ß represents the item parameters of the item, and $Q(\theta) = 1 - P(\theta)$.

In stated form, this equation is the probability of an examinee's score pattern, given that individual's proficiency level and the parameters of the items answered is equal to the product of the probabilities generated from the item response model for each of the items, or the ICCs. The first term of the equation reflects the ICC for correct responses and the second term reflects those for incorrect responses. The result of this multiplication of the ICCs is known as the posterior distribution of proficiency. The maximum likelihood estimate method is merely the mode of this posterior distribution or probability. Other types of proficiency estimation methods are variations of this idea. The Bayes Modal Estimate utilizes information of the proficiency level prior to the

observation of the score pattern and is commonly called the prior distribution of θ. This information is treated as one more item added into the overall estimation scheme. The width of the posterior distribution of proficiency is often used as a measure of the estimate accuracy. The narrower the distribution, the more accurate the estimate of proficiency is. It can be shown that lengthening a test with items of appropriate difficulty will narrow the distribution and provide increased precision (Wainer & Mislevy, 2000). Certainly, accuracy of proficiency estimation is crucial for high stakes examinations such as licensure tests. Thus another area of ongoing research is the development of alternative procedures for proficiency level estimation to enhance accuracy such as weighted maximum likelihood estimation (WLE), expected a posteriori estimation (EAP), and maximum a posteriori estimation (MAP) (Meijer & Nering, 1999).

Cheng and Liou (2000) provide a review of several methods currently used to estimate proficiency levels in CATs and examined the accuracy of different testing algorithms (discussed below) through the evaluation of mean squared errors (MSEs) associated with the resulting proficiency level estimate. Noting that MLE and Bayesian modal estimations of have been shown to give biased estimates in short tests, they investigated combinations of MLE and WLE with four methods of item selection; optimal item difficulty, most informative item, and two versions of Kullback-Leiber (KL) information. Their results suggested that in all combinations correction of proficiency estimate bias was necessary in the earlier stages of CAT. WLE gave less biased estimates regardless of the algorithm used. While the WLE and KL item selection combination outperformed the other combinations, having the smallest amount of MSE throughout the simulation, it was more time consuming, requiring an average of .25 seconds to select an

item. Given the similarities in overall results, they recommended considering the use of WLE combined with optimal item difficulty for CATs with larger banks or longer tests and the WLE/KL combination for smaller banks or when time limitations are not an issue. The effects of other parameters such as content or item exposure still need to be investigated, however, as they may effect these recommendations.

Testing Algorithms: Item Selection and Exposure Rate

The rules specifying the questions to be answered by the examinee and their order of presentation is known as a CAT testing algorithm or item selection rule (ISR). These rules establish how to start the examination, how to continue and when to stop. As previously noted, successful implementation of CAT depends upon the efficiency and usefulness of the item selection criterion. The most widely used item selection strategies are the maximum information approach and Bayesian item selection, developed by Lord and Owen respectively (Cheng & Liou, 2000; Thissen and Mislevy, 2000). It is well established that both these methods converge on true proficiency level, yet the speed at which a particular algorithm obtains this result depends on the initialization of the process. Van der Linden (1999) suggests two reasons why it is desirable for the algorithm to begin with estimations of the proficiency level close to its true value. Knowing that optimal content validity is paramount in real world administration and that this requires additional constraints to be placed on item selection that in turns slows the algorithm, precise initial estimation of the proficiency level would alleviate some of the problem. Additionally concerns about item exposure rates may be decreased with more accurate initial estimates of theta.

Chang and Ying (1999) detail another method for item selection based on the discrimination parameter of the items. They reasoned that it might be advantageous to utilize the more highly discriminating items later in the testing algorithm when the estimates of proficiency were more refined as a result of increased number of item responses. They noted that while the Sympson-Hetter method (which includes an exposure parameter) effectively controls exposure rate of all items, it does little to increase the exposure rate of those items rarely selected. Such an increase could enhance the efficiency of the overall item pool. Their method, *a*- stratified multistage CAT, separates the items in the pool into a number of levels based on their *a* (discrimination) levels. During the first parts of the examination, the testing algorithm selects those items with lower slopes while those with higher levels are utilized later. At each level, an optimization criterion that matched the *b* (difficulty) to the estimated proficiency level is used in the selection process rather than maximum item information approach. Simulation studies comparing this method to CAT utilizing maximum information approach, Bayesian selection and the Sympson-Hetter method of item exposure control indicated the *a*-stratified multistage approach produced lower than average exposure of the other methods. Additional study to investigate the effects of item bank size, number of examinees, content balancing and variable length of tests is suggested by the authors. Chen, Ankenmann, and Chang (2000) compared five types of item selection rules that are either currently in large use or have been proposed as alternatives. Their study is one of the firsts to examine the effects of the various selection rules on efficiency and precision of proficiency estimation. Although beyond the scope of this work, readers with strong methodological backgrounds may find this particular article helpful.

Although CAT provides a distinct test for each examinee, test pool items may be used for more than one individual as selected according to the testing algorithm's choice of the "best" item to present. Inflated scores of subsequent examines may result as items become known as a result of frequent administration or item exposure rate. Certainly, the higher the stakes of the examination, the more likely that individuals will attempt to gain information about test pool items. Additionally, breadth and depth of the item bank will contribute to exposure rates. In recent years, considerable attention has been afforded to the risk of test security as a result of high item exposure rates. Several methods have been proposed to control the rates to include management of item banks, examination, correction of unusual responses following test administration, and a variety of procedures to control exposure rates during the test administration. (Chang & Twu, 1998; Meijer & Nering, 1999; Thissen & Mislevy, 2000). The issue of interest for this paper will be that of exposure rate control methodology.

Chang and Twu (1998) compared five leading algorithms proposed to control item exposure rates in terms of test security, item overlap rate, and the conditional standard error of measurement. The simplest of these, developed by McBride and Martin, chooses the first test item randomly from the top five choices, the second from the top four, and so on until the fifth item at which point the best available item is presented. The Sympson-Hetter procedure determines an exposure control parameter for each item in the pool. The decision to present the item depends upon the value of this parameter. High use items usually carry low parameter values while items rarely used may have parameter values approaching 1.0, and are thus almost always administered if selected by the testing algorithm. The Davey-Parshall method also utilizes an exposure control parameter,

however, this parameter is conditioned on all other items administered previously. Additionally, this method purports to minimize the extent to which sets or pairs of items are presented in the examination. The method requires the use of an exposure table developed through a series of simulation studies. The Stocking-Lewis unconditional multinomial procedure is a remodeled approach of the Sympson-Hetter described above. The exposure control parameter is developed in the same manner but the next item selection process employs a multinomial model rather than the use optimal item. Stocking-Lewis also developed the conditional multinomial method for use in controlling exposure rates to examinees with the same or similar levels of proficiency. Chang and Twu (1998) concluded that the McBride-Martin method, though simple, did not ensure item security any better than testing completed without exposure controls in place. This finding suggests that controlling item exposure rates early in the examination does not remedy the test security issue. The Sympson-Hetter as well as the Stocking-Lewis unconditional multinomial procedures were found to yield similar results on all criteria investigated, however, the development of the parameters was more efficiently accomplished with the Sympson-Hetter. Both the Davey-Parshall method and the Stocking-Lewis conditional procedure controlled item exposure rates and item overlap rates but at the price of loss in measurement precision. The Davey-Parshall method loss was deemed acceptable in light of the test security it afforded. It is yet to be determined if the most satisfactory results of Stocking-Lewis are worth the cost in terms of standard error of measurement.

The utilization of CAT in the assessment is not without its limitations and concerns. Researchers in attempting to produce more reliable and valid measurement

must continue their dedication to accurate definition of traits and their related domains, to the development of items measuring those traits that withstand field testing, and the construction of tests subjected to confirmatory studies. The 21st century will no doubt see the expansion of IRT understanding and its implications for CAT as much remains to be elucidated. Limitations of time and space have precluded the introduction or expansion of issues concerning pool development such as bank size, item quality, pool integrity, invariance drift, and presence of multidimensionality; issues concerning administration and scoring such as content balancing, test speed, item review, and equating methodologies; and certainly, issues concerning the examinees. Individuals involved in assessment must remain committed to enhancing their own knowledge base on these issues and it is hoped that this work provides a starting point for that endeavor. (Hambleton, 2000; Hays, et al., 2000)

References

Chang, H., & Ying, Z. (1999). *a*-Stratified multistage computerized adaptive testing. *Applied Psychological Measurement, 23*(3), 211-222.

Chang, S., & Twu, B. (1998). *A comparative study of item exposure control methods in computerized adaptive testing.* ACT Research Report Series 98-3.

Chen, S., Ankenmann, R. D., & Chang, H. (2000). A comparison of item selection rules at the early stages of computerized adaptive testing. *Applied Psychological Measurement, 24*(3), 241-255.

Cheng, P.E., & Liou, M. (2000). Estimation of trait level in computerized adaptive testing. *Applied Psychological Measurement, 24*(3), 257-265.

Crocker, L., & Algina, J. (1986). *Introduction to classical and modern test theory.* Fort Worth, TX: Holt, Rinehart, and Winston.

Fan, X. (1998). Item response theory and classical test theory: An empirical comparison of their item/person statistics. *Educational and Psychological Measurement, 58*(3), 357-381.

Flaugher, R. (2000). Item pools. In H. Wainer, *Computerized Adaptive Testing: A Primer* (pp. 37-58). Mahwah, NJ: Erlbaum.

Hambleton, R. K. (2000). Advances in performance assessment methodology. *Applied Psychological Measurement, 24*(4), 291-293.

Hambleton, R. K. (2000). Emergence of item response modeling in instrument development and data analysis. *Medical Care, 38*(9), (Suppl. II), II-60-II-65.

Harvey, R. J., & Hammer, A. L. (1999). Item response theory. *The Counseling Psychologist, 27*(3), 353-383.

Hays, R.D., Morales, L.S., & Reise, S. P. (2000). Item response theory and health outcomes measurement in the 21st century. *Medical Care, 38*(9), (Suppl. II), II-28-II-42.

Meijer, R. R., & Nering, M. L. (1999). Computerized adaptive testing: Overview and introduction. *Applied Psychological Measurement, 23*(3), 187-194.

Overton, R. C., & Harms, H. J. (1997). Adapting to adaptive testing. *Personnel Psychology, 50*(1), 171-185.

Thissen, D., & Mislevy, R. J. (2000). Testing algorithms. In H. Wainer, *Computerized Adaptive Testing: A Primer* (pp. 37-58). Mahwah, NJ: Erlbaum.

Van der Linden, W. J. (1999). Empirical Initialization of the trait estimator in adaptive testing. *Applied Psychological Measurement, 23*(1), 21-29.

Wainer, H. (1983). On item response theory and computerized adaptive tests: The coming technological revolution in testing. *The Journal of College Admissions, 28*(4), 9-16.

Wainer, H. & Mislevy R. J. (2000). Item response theory, item calibration, and proficiency estimation. In H. Wainer, *Computerized adaptive testing: A primer* (pp. 61-99). Mahwah, NJ: Erlbaum.

Ware, J.E., Bjorner, J. B., & Kosinski, M. (2000). Practical implications of item response theory and computerized adaptive testing. *Medical Care, 38*(9), (Suppl. II), II-73-II-82.

Wise, S. L. (1997). *Overview of practical issues in a CAT program*. Paper presented at the meeting of the National Council on Measurement in Education, Chicago, IL. (ERIC Document Reproduction Service No. ED408330)

U.S. Department of Education
Office of Educational Research and Improvement (OERI)
National Library of Education (NLE)
Educational Resources Information Center (ERIC)

REPRODUCTION RELEASE

(Specific Document)

TM033701

I. DOCUMENT IDENTIFICATION:

Title: COMPUTERIZED ADAPTIVE TESTING: SOME ISSUES IN DEVELOPMENT

Author(s): VENETIA L. ORCUTT, MBA, PA-C

Corporate Source: UNIVERSITY OF NORTH TEXAS

Publication Date: 12/2001

II. REPRODUCTION RELEASE:

In order to disseminate as widely as possible timely and significant materials of interest to the educational community, documents announced in the monthly abstract journal of the ERIC system, *Resources in Education* (RIE), are usually made available to users in microfiche, reproduced paper copy, and electronic media, and sold through the ERIC Document Reproduction Service (EDRS). Credit is given to the source of each document, and, if reproduction release is granted, one of the following notices is affixed to the document.

If permission is granted to reproduce and disseminate the identified document, please CHECK ONE of the following three options and sign at the bottom of the page.

The sample sticker shown below will be affixed to all Level 1 documents

PERMISSION TO REPRODUCE AND DISSEMINATE THIS MATERIAL HAS BEEN GRANTED BY

Sample

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

1

Level 1

☑

Check here for Level 1 release, permitting reproduction and dissemination in microfiche or other ERIC archival media (e.g., electronic) *and* paper copy.

The sample sticker shown below will be affixed to all Level 2A documents

PERMISSION TO REPRODUCE AND DISSEMINATE THIS MATERIAL IN MICROFICHE, AND IN ELECTRONIC MEDIA FOR ERIC COLLECTION SUBSCRIBERS ONLY, HAS BEEN GRANTED BY

Sample

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

2A

Level 2A

☐

Check here for Level 2A release, permitting reproduction and dissemination in microfiche and in electronic media for ERIC archival collection subscribers only

The sample sticker shown below will be affixed to all Level 2B documents

PERMISSION TO REPRODUCE AND DISSEMINATE THIS MATERIAL IN MICROFICHE ONLY HAS BEEN GRANTED BY

Sample

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

2B

Level 2B

☐

Check here for Level 2B release, permitting reproduction and dissemination in microfiche only

Documents will be processed as indicated provided reproduction quality permits.
If permission to reproduce is granted, but no box is checked, documents will be processed at Level 1.

I hereby grant to the Educational Resources Information Center (ERIC) nonexclusive permission to reproduce and disseminate this document as indicated above. Reproduction from the ERIC microfiche or electronic media by persons other than ERIC employees and its system contractors requires permission from the copyright holder. Exception is made for non-profit reproduction by libraries and other service agencies to satisfy information needs of educators in response to discrete inquiries.

Sign here, please

Signature: Venetia L Orcutt	Printed Name/Position/Title: VENETIA L. ORCUTT, MBA, PA-C
Organization/Address: UT SOUTHWESTERN MED CENTER 5323 HARRY HINES BLVD DALLAS TX 75390-9090	Telephone: 214 648 1701 FAX: 214 648 1003
	E-Mail Address: VENETIA.ORCUTT@UTSOUTHWESTERN.EDU Date: 2/15/2002

(Over)

III. DOCUMENT AVAILABILITY INFORMATION (FROM NON-ERIC SOURCE):

If permission to reproduce is not granted to ERIC, or, if you wish ERIC to cite the availability of the document from another source, please provide the following information regarding the availability of the document. (ERIC will not announce a document unless it is publicly available, and a dependable source can be specified. Contributors should also be aware that ERIC selection criteria are significantly more stringent for documents that cannot be made available through EDRS.)

Publisher/Distributor:
Address:
Price:

IV. REFERRAL OF ERIC TO COPYRIGHT/REPRODUCTION RIGHTS HOLDER:

If the right to grant this reproduction release is held by someone other than the addressee, please provide the appropriate name and address:

Name:
Address:

V. WHERE TO SEND THIS FORM:

Send this form to the following ERIC Clearinghouse:

However, if solicited by the ERIC Facility, or if making an unsolicited contribution to ERIC, return this form (and the document being contributed) to:

ERIC Processing and Reference Facility
4483-A Forbes Boulevard
Lanham, Maryland 20706

Telephone: 301-552-4200
Toll Free: 800-799-3742
FAX: 301-552-4700
e-mail: ericfac@inet.ed.gov
WWW: http://ericfacility.org